Explainability: An Aesthetic for Aesthetics in Computational Creative Systems

Paul M. Bodily and Dan Ventura
Computer Science Department
Brigham Young University
Provo, UT 84602 USA
paulmbodily@cs.byu.edu,ventura@cs.byu.edu

Abstract

Of continued interest in the field of Computational Creativity (CC) is the question of what characteristics are required for autonomous creativity. Many characteristics have been proposed including the possession of an autonomous aesthetic. Paramount to the idea of an autonomous aesthetic is the need for a meta-aesthetic: an aesthetic which guides the system in selecting its own aesthetic. We review how aesthetics have (and have not) been used in CC systems to date, including examples of autonomous aesthetics. We formalize the idea of a meta-aesthetic in an extension of Wiggins' 2006 framework for describing computational systems generally. We propose *explainability* as an effective meta-aesthetic for autonomous creative systems and make some comments about the explainability of creativity and of explainability itself.

Introduction

Computational creativity (CC) has been characterized as the quest for "computational systems which, by taking on particular responsibilities, exhibit behaviours that unbiased observers would deem to be creative" (Colton and Wiggins 2012). Since the dawn of CC, researchers have continually hypothesized about what, in addition to output alone, is required to create the perception of creativity in a computational system, suggesting such elements as creative processes (Ritchie 2007; Colton 2008), self-evaluation (Wiggins 2006; Jennings 2010), intention (Ventura 2016; Jordanous and Keller 2016; Guckelsberger, Salge, and Colton 2017), and self-awareness (Linkola et al. 2017). These aspects of a creative system, which often imitate characteristics of human creativity, lead the observer to sense that the system has a consciousness which impels its creative behavior.

The high goal of conscious creativity hinges to a large extent on the ability of a system to create its own *aesthetic*. Deriving from the Greek word *aisthetikos* meaning "sentient" or "perceptive", this term was coined in Alexander Baumgarten's 1735 dissertation *Meditationes philosophicae de nonnullis ad poema pertinentibus* (Philosophical considerations of some matters pertaining to the poem) to describe art as a means of knowing. An aesthetic describes a philosophy of art, a theory of criticism, a set of values or beliefs about what is beautiful and good (Mothersill 2004).

Among the definitions for *aesthetic* listed by Koren in his book *Which "aesthetics" do you mean?: Ten definitions* we are interested in those that define an aesthetic as: a "cognitive mode" or awareness of abstract and/or particular "sensory and emotive qualities"; an "opinion, belief, or attitude related to some of the underlying principles of art" which, if not explicit, can be inferred from artifacts; a "style" or "perceptually cohesive organization of qualities...that is distinct from other perceptually cohesive organizations of qualities"; and an "ability to make judgments of value" (2010). In short, we think of an aesthetic as an opinion, belief, or attitude about principles of art (in the broadest sense of the term) of which there is some cognitive awareness and which serves to make judgments of value related to style.

Some have attempted to distinguish between the evaluation of creativity (with emphasis on novelty) and aesthetics (with emphasis on pleasure or beauty) (Cohen et al. 2012). This distinction is valid when *aesthetics* is used to refer to "superficial appearance" or as a synonym for *taste* or *beauty* (Koren 2010); however, in the definition we adopt an *aesthetic* encompasses all qualities that are necessary to judge a piece of art as successful or, in our case, creative.

It has also been noted that in considering aesthetics as an "evaluative discipline," there is a distinction between judgment and evaluation (Cohen et al. 2012): whereas *evaluations* represent technical, objective, quantifiable *measures* (e.g., 37° C), *judgments* represent "human", subjective, qualitative *values* (e.g., "it's hot"). This distinction poses a critical challenge for discussing aesthetics in computational systems (which by nature lack human subjectivity) and raises questions about whether qualitative judgments merely represent or can somehow be represented by complex quantitative evaluations. This discussion is beyond our scope, but for the purposes of this paper we assume that an aesthetic judgment can be represented as a quantitative function.

The idea of incorporating an aesthetic into a computational system has been broadly discussed and many CC systems (implicitly or explicitly) define aesthetics. So also has the topic of initiating and changing a system's aesthetic been addressed in various places (e.g., (Jennings 2010)). While many have advocated the enhanced creativity of a system which possesses its own aesthetic, we find that little has been said about what principles can be used to guide the system

in selecting of a "good" aesthetic. In short what is needed is an aesthetic for aesthetics.

In what follows we review what has been said elsewhere about aesthetics and autonomous aesthetics as they relate to computational systems. We also review what aesthetics have been proposed for CC systems. We finally turn to the idea of a *meta-aesthetic*, or an aesthetic for evaluating aesthetics, and propose *explainability* as one such meta-aesthetic. Our purpose is two-fold: first, to bring attention and add fuel to the assertion that a system that is aesthetically autonomous is more creative than one that is not; and second, to argue that because creativity is a fundamentally social construct, explainability is a critical characteristic of a creative system's meta-aesthetic.

Background: Aesthetics in CC Systems

The concept of an aesthetic (as we have defined it above) has been referenced using a variety of terms in the CC literature. Papadopoulos and Wiggins lament that "the big disadvantage of most, if not all, the computational models (in varying degrees) is that... the computers do not have *feelings*" (emphasis added) (1999). Boden, in her seminal work on creativity and artificial intelligence, talks of a "pre-existing mental structure" or " hidden mental faculty which has positive evaluation built in to it" (2004). In describing a framework of crucial properties for creative systems, Wiggins concretizes this built-in "positive evaluation" as a "*set of rules*" $\mathscr{E}$ for evaluating concepts "according to whatever criteria we may consider appropriate" (2006). Jennings describes autonomous systems as possessing the ability to initiate and guide changes to its "standards" and generate its own "opinions" (2010). Each of these terms highlight aspects that we have previously identified as defining an aesthetic.

Attempts at formalizing computational aesthetics span nearly a century. George David Birkhoff is credited with having fathered computational aesthetics in his 1933 book *Aesthetic Measure* in which he defines aesthetic measure as the ratio of *order* to *complexity*. From this definition came other aesthetic measurements including Shannons entropy (Shannon 2001) and Kolmogorov complexity (Rigau, Feixas, and Sbert 2007). When applied to concrete aspects of an artefact these measures represent aesthetics evaluations that have been used in many computational systems.

On the assumption that an aesthetic encompasses all qualities that are necessary to judge an artefact as creative, two common aesthetic qualities used in the judgment of CC artefacts are novelty and typicality (Ritchie 2007). One way that these qualities have been measured is the Wundt curve (Berlyne 1970) (see Figure 1). The curve represents the value of an artefact as novelty increases. Initially value increases as new ideas and features are incorporated into the artefact. At some point, however, the artefact becomes so new that it begins to no longer fit within the domain of interest and the value decreases until it is no longer of interest.

In addition to *novelty* and *typicality*, several other aesthetic values have been presented in the CC literature including: *skill*, *imagination*, and *appreciation* (Colton 2008),

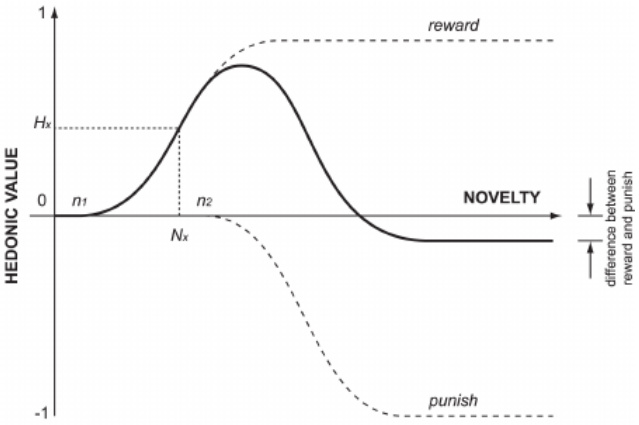

Figure 1: The Wundt curve represents an example of a computational aesthetic quality commonly used for evaluating novelty and typicality. As the novelty of an artefact increases, so *to an extent* does the value. At some point the artefact strays far enough from the bounds of "normal" to prevent observers from recognizing even limited value. Figure from Saunders et al. (2010).

as well as *value* and *surprise* (Boden 2004). Hofstadter suggests that *complexity* is an aesthetic for creativity (1980).

Every creative system is reflective of an aesthetic (Koren 2010), and in some cases the aesthetic is explicitly modeled. CC systems have used a number of different aesthetic implementations including neural networks trained on user ratings (Morris et al. 2012; Monteith, Martinez, and Ventura 2010; Heath and Ventura 2016); Markov models (Abgaz et al. 2017); and "baked-in" knowledge (Ventura 2016) and many others.

It is a common assumption that possessing an autonomous aesthetic, independent from that of the programmer or designer, is a fundamental characteristic of creative systems. Colton (2008) asserts that ultimately the perception of creativity will require that CC systems "develop their own aesthetic along with their own styles". Guckelsberger, Salge, and Colton propose moving away from anthropocentric models and present the *enactive artificial intelligence* framework as a minimal model of intentional creative agency (2017).

Systems can be found that to varying extents develop an autonomous aesthetic. Cook and Colton (2015) present a painting system that evolves its own preference functions which enable it to make non-random, consistent aesthetic choices that are not based on any external, existing opinion. DARCI implements separate semantic and image generation models and uses the semantic model autonomously to guide the process of rendering images that convey particular concepts, including those not seen in training (Heath and Ventura 2016). Jennings (2010) argues that a creative system must not only have an autonomous aesthetic, but must also be capable of autonomously *changing* its own aesthetic in non-random ways (see also Ackerman et al. (2017)).

A Framework for Describing Aesthetics

As several in the field have asserted (and which we also assert), an autonomous aesthetic is a key component of creative systems. This assertion naturally begs the question: What are the characteristics that a computational system might use for *selecting* an aesthetic?

To more precisely consider this question, we present a framework for describing, analyzing, and comparing aesthetics. The framework might be thought of as being analogous and a possible extension to the framework of Wiggins (2006), but with focus on aesthetics instead of concepts. In Wiggins' framework for describing, analyzing, and comparing concepts, a set of rules $\mathscr{E}$ for evaluating concepts is presented as a crucial component of creative systems. The function computed by $\mathscr{E}$—denoted by Wiggins as $[[\mathscr{E}]]$—yields a real numbered value in the range [0,1] representing the distribution of evaluation scores $[[\mathscr{E}]](c)$ over all concepts c in the concept universe. While Wiggins chooses to forgo any discussion on what is included in $\mathscr{E}$, we make the simplifying assumption (similar to (Jennings 2010)) that $[[\mathscr{E}]]$ is equivalent to a system's standards or aesthetic (we later discuss the validity of this assumption). It is also significant to our discussion to note that a language, $\mathscr{L}$, is needed in order to express this set of rules $\mathscr{E}$ (Wiggins 2006).

Let us hypothesize that there exists an aesthetic universe $\mathscr{A}$ which encompasses all possible aesthetics.

Definition 1 (*Aesthetic universe*). The *aesthetic universe*, $\mathscr{A}$, is a multidimensional space, whose dimensions allow for the representation of any aesthetic, and all possible distinct aesthetics correspond with distinct points in $\mathscr{A}$.

A more precise formulation of an aesthetic a is beyond the scope of this paper and when discussing those aesthetics included in $\mathscr{A}$ we include abstract and concrete, partial and complete aesthetics.

Axiom 1 (*Aesthetic universality*). All possible aesthetics are represented in $\mathscr{A}$; thus, $\mathscr{A}$ is the type of all possible aesthetics.

Axiom 2 (*Non-identity of aesthetics*). All aesthetics, a_i, represented in $\mathscr{A}$ are mutually non-identical. $\forall a_1, a_2 \in \mathscr{A}, a_1 \neq a_2$.

We now define a set of rules, $\mathscr{T}_{\mathscr{A}}$, which allows a traversal of $\mathscr{A}$ according to some search strategy and a set of rules, $\mathscr{E}_{\mathscr{A}}$, for evaluating the quality of any aesthetic found. $\mathscr{A}$ can be enumerated using an interpreter $\langle\langle .,. \rangle\rangle$, which, given $\mathscr{T}_{\mathscr{A}}$ and $\mathscr{E}_{\mathscr{A}}$, maps an ordered subset of $\mathscr{A}$, a_{in}, to another ordered subset of $\mathscr{A}$, a_{out}:

$$a_{out} = \langle\langle \mathscr{T}_{\mathscr{A}}, \mathscr{E}_{\mathscr{A}} \rangle\rangle (a_{in}).$$

In essence, $\mathscr{T}_{\mathscr{A}}$ might be thought of as a strategy for mutating a system's aesthetic as a function of previous aesthetics and $\mathscr{E}_{\mathscr{A}}$ as an evaluation mechanism for aesthetics. It is beyond our scope to consider what $\mathscr{T}_{\mathscr{A}}$ might look like, though others have devoted some significant thought to this idea (e.g., Jennings(2010)).

The discussion on which we choose to focus revolves instead around the set of rules for evaluating aesthetics, $\mathscr{E}_{\mathscr{A}}$, and the language, $\mathscr{L}$, which is used to express a particular aesthetic, a_i.

Explainability: An Aesthetic for Aesthetics

What makes a good aesthetic? As evidenced by the variety of aesthetics implemented (implicitly or explicity) in extant CC systems, many have pondered what good aesthetics for CC systems might be. *Our* purpose is rather to consider what all "good" aesthetics have in common.

Our proposal for what makes a good meta-aesthetic hinges on the idea that *creativity is an inherently social construct*. In his book on creativity Csikszentmihalyi writes: "Creativity does not happen inside people's heads, but in the interaction between a person's thoughts and a sociocultural context" (1996). He also comments: "Over and over again, the importance of seeing people, hearing people, exchanging ideas, and getting to know another person's work and mind are stressed by creative individuals."

To push this point further, let us consider for a moment a slightly adapted version of Colton's allegory of the *'Dots 2008'* exhibit (2008). In the original allegory, two painters display paintings composed of a "seemingly random arrangement of dots of paint". The story goes that despite the fact that they appear identical, an observer falls in love with one painting when its artist explains that unlike his colleague, whose painting represents nothing more than randomly placed dots, in *his* painting "each dot represents a friend of mine. The colour of the dot represents how I feel about them, and the position indicates how close I am to them." Colton uses this allegory to illustrate that creativity lies as much in the process as in the output.

To adapt the allegory, consider now that both paintings were inspired to represent the painters' feelings toward their friends. In this version of the story, the art-lover then asks each painter: "What made you decide to paint your feelings towards your friends?" The first painter responds: "My friends are important to me." The second shrugs, gestures towards the first, and responds: "My art teacher told me to." Returning a week later with a friend, the art-lover explains to the friend that both paintings represent the creativity of the first artist.

The original version of the allegory was used by Colton to demonstrate that the perception of creativity depends as much on the explanation of the creative process as it does on the artefact itself. The *extended* version of the allegory emphasizes that the perception of creativity depends equally as much on explanation of the *aesthetic*.

This line of thought leads us to the proposition of *explainability* as an aesthetic for aesthetics, that is, the idea that a good aesthetic can be *explained*. In support of this idea, it is interesting to note that in some contexts the term *aesthetic* is even defined as "the verbal or written statement itself" of beliefs about art (Koren 2010).

To our knowledge very few systems—and no systems with an autonomous aesthetic—exist which attempt to ex-

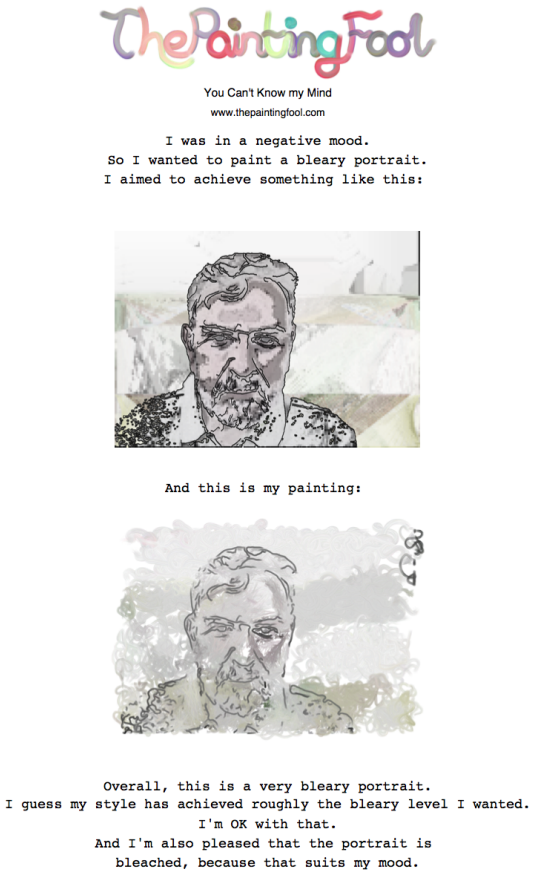

Figure 2: In its *You Can't Know my Mind* exhibit, The Painting Fool paints portraits reflective of a "mood" and possibly an aesthetic. Figure from Colton and Ventura (2014).

plain their aesthetic, though it is not uncommon to see presuppositions about a "system [that] has unlimited capacities to enter into a dialogue and to frame its actions" (2017). Two notable exceptions to the dearth of systems that explain their aesthetic are the *You Can't Know my Mind* exhibit from The Painting Fool (Colton and Ventura 2014) and the latest version of DARCI's image creation process (Heath and Ventura 2016).

Colton and Ventura's *You Can't Know my Mind* exhibit features The Painting Fool as it paints (or chooses not to paint) portraits that reflect its current mood and aesthetic (see Figure 2). In addition to displaying the painting, it explains some of what the system was "thinking" and "feeling" as it painted the portrait as well as how well it felt (with the help of DARCI) that it accomplished its intention. Though the system does reflect and explain an autonomous mood and interpretation of particular descriptors, this explanation is arguably not representative of an aesthetic because the system is explaining more its observations and logic rather than its system of values.

DARCI (Heath and Ventura 2016) is another example of a system that approaches the threshold of explaining its aesthetic (see Figure 3). The system begins with an inspiring image from which it tries to "think" of similar looking objects. It then creates a new image of a similar looking object that has been stylistically modified to reflect an aesthetic quality similar to the original. Its explanation of this process, like that of The Painting Fool, focuses primarily on describing the logical process of creating the image, but does give an impression of having consciously thought about aesthetic qualities in its creativity.

These two examples suffice to demonstrate that an explainable aesthetic *can* contribute to the perception of a sentient, aesthetically-driven creative system. We await future work to provide corroborative empirical evidence.

Enabling a system to explain its own aesthetic adds a potentially significant degree of effort. But its importance cannot be overstated. Like an overly involved parent, a CC researcher that does not equip a system with the ability to explain its own aesthetic (and is rather constantly butting in to do the explaining themselves) creates a crutch for their systems, never fully realizing the lack of creativity in their offspring until it is left to flounder on its own.

An explainable aesthetic is notably different from an explained aesthetic. Often humans participate in creativity without ever explaining their aesthetics. But this is different than assuming that their aesthetics are not explainable. Many, including Boden (2004), have argued that humans can be creative without being able to explain the aesthetic that motivates their creativity. We would argue, however, that for any human aesthetic *some* degree of explainability exists, even if it be as unconventional as "random" or "anti-aesthetic". We discuss this more below.

The purpose of an aesthetic is to impose *value*. An *explainable* aesthetic makes it possible for a creator to communicate this value to others. An *explained* aesthetic makes it possible for others to understand *why* an artifact is valued and possibly, then, to appreciate it (more) themselves.

A Language for Explaining Autonomous Aesthetics

Explainability of an aesthetic requires a language in which to express the aesthetic. To introduce what that language might be, consider that Wiggins defines a language $\mathcal{L}$ for representing $\mathcal{E}$. How does $\mathcal{E}$ differ from the aesthetic represented by $[[\mathcal{E}]]$? The answer lies in the fact that there could be many rule sets $\mathcal{E}_i$ which describe the *same* function (i.e., domain, co-domain, and range are equal). Considering that each of these rule sets is explained using a language $\mathcal{L}$, this essentially means that there could be many explanations of an aesthetic that are functionally equivalent. How then does a CC system decide which explanation is best? Given two distinct rule sets $\mathcal{E}_1$ and $\mathcal{E}_2$, let $\mathcal{E}_1 \sim \mathcal{E}_2$ mean that $[[\mathcal{E}_1]] = [[\mathcal{E}_2]]$. If $\mathcal{E}_1 \neq \mathcal{E}_2$ but $\mathcal{E}_1 \sim \mathcal{E}_2$, which explanation is to be preferred?

One likely suggestion is to consider the relationship between the amount of information contained within the explanation and the length of the explanation. The Kolmogorov complexity K, which is related to Shannon's entropy H but is *language-aware*,[1] provides just such a measure and has

[1]Note that there is an important subtlety here. There is a description language $\mathcal{D}$ associated with the definition of K. This should *not* be confused with the explanation language $\mathcal{L}$.

(a) source (b) example (c) sketch (d) rendered (e) **Overdress**

Figure 3: An example showing the intermediate images of each step of DARCI's image creation process. A generated description (personifying the system) is as follows: "I was looking for inspiration from this image (a), And it made me feel **gloomy** and **dreamy**. It also made me think of this image that I've previously seen (b), which is a picture of a **poncho**. So I started an initial image of my own by searching for a background image on the Internet based on **poncho**, **gloomy**, and **dreamy**. Then I took basic iconic images associated with those concepts and resized/placed them on the background according to how relevant they were. This was the result (c). I then modified it in a style related to **poncho**, **gloomy**, and **dreamy**, which resulted in this image (d). I did a final modification based on aesthetic quality and how closely the style related to the original image (e). The end result perhaps looks more like a **cloak** or a **vestment**, and it feels particularly **gloomy**. I call it **Overdress**."

been suggested in this capacity before (Rigau, Feixas, and Sbert 2007). The Kolmogorov complexity $K(\mathscr{E})$ of an explanation $\mathscr{E}$ of an aesthetic is distinct from the complexity of the aesthetic $[[\mathscr{E}]]$ itself (which might be measured using entropy H). Returning to the equivalent explanations $\mathscr{E}_1, \mathscr{E}_2$, this means in general that because $\mathscr{E}_1 \sim \mathscr{E}_2$, $H(\mathscr{E}_1) = H(\mathscr{E}_2)$, but because $\mathscr{E}_1 \neq \mathscr{E}_2$, $K(\mathscr{E}_1) \neq (\mathscr{E}_2)$, and the explanation with lower complexity is to be preferred.

K and H, however, are related through $\mathscr{L}$—while H is language-agnostic, K is not. If $\mathscr{E}_1$ and $\mathscr{E}_2$ are expressed in the same language, this is perhaps not noteworthy; however, if they are expressed in distinct languages $\mathscr{L}_1$ and $\mathscr{L}_2$, things become more interesting. If we think of the set of all possible aesthetics $\mathscr{A}$ (i.e., the set of all possible $[[\mathscr{E}]]$ or the set of all possible distributions over the concept universe), then depending on the choice of $\mathscr{L}$, aesthetics of high or low entropy may be considered good. In other words, the complexity of the aesthetic $H([[\mathscr{E}]])$ is not correlated with its goodness; rather, the complexity of its *explanation* $K(\mathscr{E})$ is, and that is language dependent.

This suggests at least two other ideas:

- The invention of new languages $\mathscr{L}$ is as interesting as the invention of new aesthetics $[[\mathscr{E}]]$ (as touched on by (Saunders and Grace 2008)).
- The discovery of new $\mathscr{E}$ is not the same as the discovery of new $[[\mathscr{E}]]$. It is still useful though, as reductions in $K(\mathscr{E})$ signify improved communicability.

The idea that particular languages encourage particular aesthetics is well-known in the realm of natural language. It has long been known that language guides cognition and that conceptual knowledge is shaped by a person's language (Whorf et al. 2012). It has also been shown, for example, that people make different choices based on whether the decision is framed in a native or foreign language (Keysar, Hayakawa, and An 2012). Words with significant cultural and aesthetic value often have no ("good") translation from one language to another. Honorifics which exist in some languages are absent in others. Even grammars themselves can be indicative of a culture's belief system.

The variable impact of language on the goodness or complexity of an explanation leads us to postulate a "no free lunch theorem of aesthetic languages" (which we do not attempt to prove).

Theorem 1 *The no free lunch theorem of aesthetic languages*. Given two languages, $\mathscr{L}_1$ and $\mathscr{L}_2$,

$$\sum_{[[\mathscr{E}]]\in\mathscr{A}} K(\mathscr{E}^*_{\mathscr{L}_1}) = \sum_{[[\mathscr{E}]]\in\mathscr{A}} K(\mathscr{E}^*_{\mathscr{L}_2})$$

where $\mathscr{E}^*_{\mathscr{L}_1}$ and $\mathscr{E}^*_{\mathscr{L}_2}$ are explanations of aesthetic $[[\mathscr{E}]]$ in languages $\mathscr{L}_1$ and $\mathscr{L}_2$, respectively, and $K(\mathscr{E}^*_{\mathscr{L}}) \leq K(\mathscr{E}'_{\mathscr{L}})$ for all $\mathscr{E}'_{\mathscr{L}} \sim \mathscr{E}^*_{\mathscr{L}}$. That is, $\mathscr{E}^*_{\mathscr{L}_1}$ ($\mathscr{E}^*_{\mathscr{L}_2}$) is the least complex explanation for aesthetic $[[\mathscr{E}]]$ in language $\mathscr{L}_1$ ($\mathscr{L}_2$).

In other words, if a language admits low complexity of explanation for some set of aesthetics $\mathscr{B} \subset \mathscr{A}$ then it necessarily pays for that with unavoidably higher complexity explanations for the set of all remaining aesthetics $\mathscr{A} \setminus \mathscr{B}$.

This leads to another important point about explainability: to the extent that the *shared* language between a system and its audience is different, its ability to share aesthetics will be limited (cf. (Saunders and Grace 2008)).

Does Wiggins' $[[\mathscr{E}]]$ compute an Aesthetic?

We have previously assumed that the function $[[\mathscr{E}]]$ computed by interpreting the rule set $\mathscr{E}$ is equivalent to some aesthetic $a \in \mathscr{A}$. We now consider the argument that the function $[[\mathscr{E}]]$ is instead somehow distinct from any aesthetic. To prove this argument we need to demonstrate an instance in which $[[\mathscr{E}]]$ differs from that of the aesthetic a. This would

suggest that there are other factors besides a which determine $[[\mathscr{E}]]$. We believe that most scenarios which might demonstrate this contradiction fit into one of two categories, which we call *domain rectification* and *aesthetic transfer*. Prior to summarizing the general characteristics of these categories, we share one thought experiment representative of each category.

History and Allegory Imagine a scholar who encounters a volume of ancient text. The scholar, who values historical accuracy, determines in the course of reading the text that the events described could not be historically accurate and dismisses the text as being of questionable value. A short time later, the scholar is informed by a friend that the volume in question was intended as an allegory rather than as an historical account. Rereading the text through this lens, she now finds value in the insights afforded by the allegorical interpretation.

It might appear in this example that the scholar's evaluation, $[[\mathscr{E}]](c)$, of the text c changed while her aesthetic, a, did not. We would argue, however, that there was no change in either $[[\mathscr{E}]](c)$ or a. Rather the text is being recategorized into a different domain resulting in a different evaluation $[[\mathscr{E}]]'$. In defining his framework, Wiggins' notably states that $\mathscr{E}$ is used to evaluate concepts *within a specific concept domain*, not the greater concept universe (2006). Seen in this light, there are in fact two different evaluation functions ($[[\mathscr{E}_{history}]]$ and $[[\mathscr{E}_{allegory}]]$) and also two different aesthetics ($a_{history}$ and $a_{allegory}$) at play here. The parallel changes from $[[\mathscr{E}_{history}]]$ to $[[\mathscr{E}_{allegory}]]$ and from $a_{history}$ to $a_{allegory}$ provide a plausible explanation that avoids the conclusion that $[[\mathscr{E}]]$ must be distinct from a.

The Moody Young Pianist Imagine a moody young pianist enrolled to study piano. Despite his interest in other forms of music, his instructor insists on teaching him classical music and assigns him to learn Brahms Rhapsody in G Minor Op. 79, No. 2. The boy at first does not like the piece, but in the course of time his instructor invites him to think of the piece as an interpretation through dynamics, tempo, and melodic expression of his own life experiences. This idea excites the boy, who develops a love and ownership for the piece he once despised.

It may seem that although the boy's evaluation $[[\mathscr{E}]](c)$ of the piece c had changed, his aesthetic a did not. Just as $[[\mathscr{E}]]$ is relative to a particular domain, so too (we will assume now and question later) is a domain-dependent. Here, we suggest that an aesthetic a_1 from one domain (e.g., self-expression) is transferred to another domain (e.g., classical music) with had been associated with aesthetic a_2; effectively, a_1 (temporarily or partially) replaces a_2. Thus the apparent change in $[[\mathscr{E}]]$ is associated with a change in aesthetic. This scenario is also plausibly explained without having to conclude that $[[\mathscr{E}]]$ is in any way distinct from a.

These thought experiments are representative of scenarios in which it may appear that the function $[[\mathscr{E}]]$ changes while the aesthetic a does not; however plausible arguments can be made in both cases that avoid this conclusion:

- **Domain rectification:** Apparent changes in $[[\mathscr{E}]]$ actually result from the application of different domain-specific evaluation functions (e.g., $[[\mathscr{E}_i]]$ is replaced by $[[\mathscr{E}_j]]$, where $\mathscr{E}_i \nsim \mathscr{E}_j$).
- **Aesthetic transfer:** Changes in $[[\mathscr{E}]]$ within a concept domain $\mathscr{C}$ occur in association with the (temporary or partial) adoption of an aesthetic a_j usually associated with some other concept domain $\mathscr{C}'$ and its being used in place of a_i for $\mathscr{C}$.

While these thought experiments do not prove our assumption of the equivalence of $[[\mathscr{E}]]$ and a, they do provide some suggestive support for the idea that $[[\mathscr{E}]] = a$. For now we will leave this an open question.

These two scenarios also serve to strengthen the argument for explainability as a meta-aesthetic. In the first scenario, the difference between the evaluations is explained by a change in the *domain* of the aesthetics. Without an explanation of the different domain-specific aesthetics (e.g., "what makes this a good historical account" versus "what makes this a good allegory"), differences in creative evaluation due to subjective opinion (which sometimes causes differences) cannot be distinguished from differences due to contrary assumptions about the contextual domain.

In the second scenario, the difference between the evaluations is explained by a *change* or expansion in the aesthetic. Without an explanation of the differential aesthetics, differences in creative evaluation due to a reversal of subjective opinion (e.g., "Brahms instinctively sounds good to me now") cannot be distinguished from those due to an expanded aesthetic ("I don't typically like Brahms, but this song has added meaning to me").

In both scenarios an explainable aesthetic is needed to allow the system to convincingly demonstrate that autonomous changes to its aesthetic are occurring in non-random ways.

The Explainability of Explainability

In proposing an aesthetic for aesthetics an interesting question arises: what happens when you evaluate the meta-aesthetic according to the meta-aesthetic? In this case the question takes on the more concrete form of how well does the aesthetic of explainability hold up under the aesthetic of explainability? How explainable is explainability?

On the one hand the concept of explainability seems readily explicable: ideas and aesthetics that can be communicated are preferred to those which can not. But when push comes to shove, there is a significant double-standard in explainability: though we ask for explainability, we rarely intend for ideas to be explained beyond a few layers of complexity. Indeed the entire discipline of epistemology exists essentially to question the explainability of explainability. Therefore, proposing explainability as a suitable meta-aesthetic is the beginning of a much larger discussion that needs to take place about what degree of explainability is conducive to a discussion of creativity.

Explainability is an interesting topic in relation to *creativity* which is so characteristically unexplainable. In fact many human observers feel that *un*explainability is a critical element of creativity: if the complete process by which an arte-

fact is created is known, then it cannot possibly be creative. Others have argued that creativity emerges when the process extends beyond some sufficient and necessary threshold of complexity (Hofstadter 1980). This is fundamentally at the heart of the debate over mere generation (Ventura 2016). In both humans and computers, too little information will not satiate the observer's desire to understand. On the other hand, too much detail can lead to tedium or (particularly in computers) an impression that the agent is purely carrying out predefined instructions (Colton 2008). Finding that balance is as much a key to the perception of creativity as it is to the discussion of explainability in general.

Discussion and Conclusion

Relevant to the discussion of explaining aesthetics in a given (possibly natural) language, it has been argued that the concept of creativity, though human-conceived, should not remain human-centric (Guckelsberger, Salge, and Colton 2017). While this may be true, it is also true that creativity does not happen in a vacuum, but emerges in the interaction between an agent's "thoughts" and a sociocultural context (Csikszentmihalyi 1996; Bown 2015; Jordanous 2015). It may not be that creativity is human-centric, but until substantial non-human sociocultural contexts are presented, it seems reasonable to expect that creativity in computational systems depends on at least an interaction with human sociocultural contexts.

Certainly, there exist many domains that will always be human-centric, yet to which it would be desirable to have CC agents contributing (e.g., medicine, drug design, autonomous vehicles, etc.) And, it is certain that in many such domains, an ability to explain process and/or product will be demanded by human "consumers".

However, even in a future in which creators and their sociocultural context are wholly non-human, we argue that the notion of explainability would remain a critical consideration, albeit possibly employing non-natural language for that explanation (e.g., see (Saunders and Grace 2008)).

We seem, to a large extent, to have focused thus far as a field on systems which (possibly by some arguably creative process) generate *artefacts*. We do so to the detriment of the field whose stated focus is on the greater umbrella of *behaviors* (Colton and Wiggins 2012). We may find benefit in increasingly promoting research toward systems which, independent of their generative abilities, are creative by virtue of their ability to interact with and internally react to their sociocultural context (see Figure 4).

An interesting open question is whether creativity is domain dependent or whether there is some abstract, core creative mechanism that is domain agnostic. The question is beyond the scope of our current treatment, but it is intimately coupled with the question of whether an aesthetic may be developed and explained independent of a particular creative domain (to which it may eventually be applied).

Systems with an ability to autonomously initiate, change, and explain an (domain-agnostic) aesthetic deriving from a sociocultural context would be a significant contribution to the field of CC, even without (domain-specific) generative capabilities. Such systems, created independently from any particular creative domain, could modularly apply themselves to various domains with compelling intentional creativity and self-evaluation. Such systems could prove useful to those who prefer to focus on more generative aspects of computational creativity by providing an out-of-the-box aesthetic model from which to derive autonomous guidance. Consider, for example, a model-view-controller system (Krasner, Pope, and others 1988) for computational creativity with an abstract model embodying an autonomously initiated and changing aesthetic which is reapplied across multiple domains through domain-specific controllers.

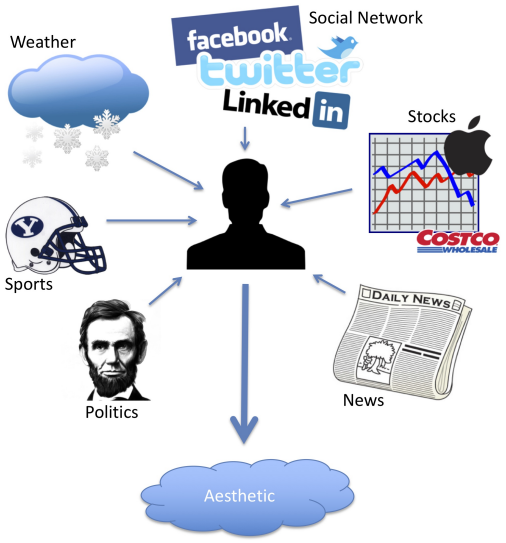

Figure 4: The ability to interact with and internally react to sociocultural context is a fundamental characteristic of creative agents. In addition to interaction with other agents, a typical human sociocultural context might include politics, sports, weather, social media, the economy, and news.

To conclude, we restate the argument of our paper using the readily available analogy of the peer review process by which this paper has been evaluated. Our manuscript is an example of an aesthetic (i.e., our opinions, beliefs, etc.) being explained by creative agents (i.e., the authors) in the demonstration of a creative artefact (i.e., the idea of explainability as a meta-aesthetic)[2]. This example itself demonstrates the thesis of our argument: *aesthetic explainability* is a minimal yet valuable standard to which we hold one another in our own creative endeavors. The success of CC systems will improve as they demonstrate similar capabilities in their own attempts to demonstrate creative behaviors.

References

Abgaz, Y.; Chaudhry, E.; ODonoghue, D.; Hurley, D.; and Zhang, J. J. 2017. Characteristics of pro-c analogies and blends between research publications. In *Proceedings of the 8th International Conference on Computational Creativity*, 1–8.

[2]This qualifies the manuscript as a meta-meta-aesthetic.

Ackerman, M.; Goel, A.; Johnson, C. G.; Jordanous, A.; León, C.; y Pérez, R. P.; Toivonen, H.; and Ventura, D. 2017. Teaching computational creativity. In *Proceedings of the 8th International Conference on Computational Creativity*, 9–16.

Berlyne, D. E. 1970. Novelty, complexity, and hedonic value. *Perception & Psychophysics* 8(5):279–286.

Birkhoff, G. D. 1933. *Aesthetic Measure*. Harvard University Press Cambridge.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. Psychology Press.

Bown, O. 2015. Attributing creative agency: Are we doing it right? In *Proceedings of the 6th International Conference on Computational Creativity*, 17–22.

Cohen, H.; Nake, F.; Brown, D. C.; Brown, P.; Galanter, P.; McCormack, J.; and dInverno, M. 2012. Evaluation of creative aesthetics. In *Computers and creativity*. Springer. 95–111.

Colton, S., and Ventura, D. 2014. You can't know my mind: A festival of computational creativity. In *Proceedings of the 5th International Conference on Computational Creativity*, 351–354.

Colton, S., and Wiggins, G. A. 2012. Computational creativity: The final frontier? In *Proceedings of the European Conference on Artificial Intelligence*, volume 12, 21–26.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *AAAI spring symposium: creative intelligent systems*.

Cook, M., and Colton, S. 2015. Generating code for expressing simple preferences: Moving on from hardcoding and randomness. In *Proceedings of the 6th International Conference on Computational Creativity*, 8–16.

Csikszentmihalyi, M. 1996. *Creativity: Flow and the Psychology of Discovery and Invention*. Harper Perennial, New York.

Guckelsberger, C.; Salge, C.; and Colton, S. 2017. Addressing the why? in computational creativity: A non-anthropocentric, minimal model of intentional creative agency. In *Proceedings of the 8th International Conference on Computational Creativity*, 128–135.

Heath, D., and Ventura, D. 2016. Creating images by learning image semantics using vector space models. In *Proceedings of the Association for the Advancement of Artificial Intelligence*, 1202–1208.

Hofstadter, D. R. 1980. *Gödel, Escher, Bach*. Vintage Books New York.

Jennings, K. E. 2010. Developing creativity: Artificial barriers in artificial intelligence. *Minds and Machines* 20(4):489–501.

Jordanous, A., and Keller, B. 2016. Modelling creativity: Identifying key components through a corpus-based approach. *PLOS One* 11(10):e0162959.

Jordanous, A. 2015. Four pppperspectives on computational creativity. In *Proceedings of the AISB Symposium on Computational Creativity*, 16–22.

Keysar, B.; Hayakawa, S. L.; and An, S. G. 2012. The foreign-language effect: Thinking in a foreign tongue reduces decision biases. *Psychological Science* 23(6):661–668.

Koren, L. 2010. *Which" aesthetics" do you mean?: Ten definitions*. Imperfect Pub.

Krasner, G. E.; Pope, S. T.; et al. 1988. A description of the model-view-controller user interface paradigm in the smalltalk-80 system. *Journal of Object Oriented programming* 1(3):26–49.

Linkola, S.; Kantosalo, A.; Männistö, T.; and Toivonen, H. 2017. Aspects of self-awareness: An anatomy of metacreative systems. In *Proceedings of the 8th International Conference on Computational Creativity*, 189–196.

Monteith, K.; Martinez, T. R.; and Ventura, D. 2010. Automatic generation of music for inducing emotive response. In *Proceedings of the 1st International Conference on Computational Creativity*, 140–149.

Morris, R. G.; Burton, S. H.; Bodily, P.; and Ventura, D. 2012. Soup over bean of pure joy: Culinary ruminations of an artificial chef. In *Proceedings of the 3rd International Conference on Computational Creativity*, 119–125.

Mothersill, M. 2004. Beauty and the critic's judgment: Remapping aesthetics. In Kivy, P., ed., *The Blackwell Guide to Aesthetics*. Blackwell Publishing Ltd. 152–166.

Papadopoulos, G., and Wiggins, G. 1999. AI methods for algorithmic composition: A survey, a critical view and future prospects. In *Proceedings of the AISB Symposium on Musical Creativity*, 110–117.

Rigau, J.; Feixas, M.; and Sbert, M. 2007. Conceptualizing Birkhoff's aesthetic measure using Shannon entropy and Kolmogorov complexity. In *Proceedings of the Third Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging*, 105–112.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17(1):67–99.

Saunders, R., and Grace, K. 2008. Towards a computational model of creative cultures. In *AAAI Spring Symposium: Creative Intelligent Systems*, 67–74.

Saunders, R.; Gemeinboeck, P.; Lombard, A.; Bourke, D.; and Kocaballi, A. B. 2010. Curious whispers: An embodied artificial creative system. In *Proceedings of the 1st International Conference on Computataional Creativity*, 100–109.

Shannon, C. E. 2001. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review* 5(1):3–55.

Ventura, D. 2016. Mere generation: Essential barometer or dated concept. In *Proceedings of the 7th International Conference on Computational Creativity*, 17–24.

Whorf, B. L.; Carroll, J. B.; Levinson, S. C.; and Lee, P. 2012. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. MIT Press.

Wiggins, G. A. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19(7):449–458.